

# A kernel-based approach to overparameterized Hammerstein system identification

Riccardo S. Risuleo, Giulio Bottegal and Håkan Hjalmarsson \*

May 19, 2016

## Abstract

The object of this paper is the identification of Hammerstein systems, which are dynamic systems consisting of a static nonlinearity and a linear time-invariant dynamic system in cascade. We assume that the nonlinear function can be described as a linear combination of  $p$  basis functions. We model the system dynamics by means of an  $np$ -dimensional vector. This vector, usually referred to as *overparameterized vector*, contains all the combinations between the nonlinearity coefficients and the first  $n$  samples of the impulse response of the linear block. The estimation of the overparameterized vector is performed with a new regularized kernel-based approach. To this end, we introduce a novel kernel tailored for overparameterized models, which yields estimates that can be uniquely decomposed as the combination of an impulse response and  $p$  coefficients of the static nonlinearity. As part of the work, we establish a clear connection between the proposed identification scheme and our recently developed nonparametric method based on the stable spline kernel.

## 1 Introduction

A nonlinear system is usually called an Hammerstein system when it is composed of two blocks in cascade, the first being a static nonlinearity and the second a linear time-invariant (LTI) dynamic system [16].

There are several areas in science and engineering where Hammerstein systems find applications, see e.g. [15], [28], [4]. For this reason, in recent years Hammerstein system identification has become a popular and rather active research topic [26], [12].

---

\*R. S. Risuleo, G. Bottegal and H. Hjalmarsson are with the ACCESS Linnaeus Center, School of Electrical Engineering, KTH Royal Institute of Technology, Sweden ([risuleo;bottegal;hjalmarsson@kth.se](mailto:risuleo;bottegal;hjalmarsson@kth.se)). This work was supported by the European Research Council under the advanced grant LEARN, contract 267381 and by the Swedish Research Council under contract 621–2009–4017.

Several approaches have been proposed for Hammerstein system identification. For instance, in [11] a kernel-based regression method is described, [10] proposes an identification approach based on a stochastic approximation, while [9] focuses on subspace methods. In [28], [3] and [24] iterative methods based on least-squares are studied.

An interesting approach was proposed by Er-Wei Bai in [2]. Here, the static nonlinearity is modeled as the linear combination of  $p$  basis functions, while the LTI system is assumed to be a finite impulse response (FIR) with  $n$  coefficients. The Hammerstein system is then modeled as a linear regression, where the regressor vector is  $np$ -dimensional. Since it contains all the combinations of the nonlinearity coefficients and the FIR coefficients, this vector is usually called *overparameterized vector*. Its estimate is obtained via least-squares and then it is decomposed in order to obtain the nonlinearity coefficients and the impulse response. Albeit proven to be asymptotically consistent, the whole procedure suffers of two main drawbacks. First, since it relies on a least-squares estimation of a possibly very high-dimensional vector, the final estimates may suffer from high variance [16]. Second, the procedure does not guarantee that the estimated  $np$ -dimensional vector can be exactly decomposed to obtain the nonlinearity coefficients and the FIR system, and thus approximations are required.

In this paper, we propose a regularization technique to curb the variance of the estimates of the overparameterized vector. Similarly to [2], we model the Hammerstein system dynamics using the aforementioned overparameterized vector, then we solve the regression problem relying on a kernel-based approach. To this end, we introduce a novel kernel, called the *Kronecker overparameterized* (KOP) kernel, which is the composition of a rank-one positive semi-definite matrix and the so-called *first-order stable spline kernel* (see [22], [21], [6], and [23] for details). The structure of this kernel depends on a few parameters (also called *hyperparameters* in this context), which we need to estimate from data. This task is addressed by an *empirical Bayes* approach [18], that is to say by maximizing the marginal likelihood (ML) of the output. Once the kernel parameters are fixed, the overparameterized vector is estimated via regularized least squares [23]. Equivalently, we can think of the overparameterized vector as a Gaussian random vector with zero-mean and covariance matrix given by the KOP kernel. With this interpretation, the estimate corresponds to the minimum mean square error estimate in the Bayesian sense [27].

A contribution of this paper is to reveal some interesting properties of the estimated overparameterized vector provided by the proposed method. We prove that, as opposed to [2], this estimate can be decomposed exactly in order to obtain the nonlinearity coefficients and the LTI system impulse response, with no loss of information due to approximations. The concept of *exact decomposition* will be made clear throughout the paper. We also demonstrate strong connections with our recently proposed method [25], effectively proving that, although the two approaches are inherently different, the estimates obtained with the two methods are equivalent. Finally, we show, through several numerical experiments, that the proposed method outperforms both the algorithm proposed in [2] and the standard MATLAB system identification toolbox function

for Hammerstein system identification.

The paper is organized as follows. In the next section, we formulate the Hammerstein system identification problem. In Section 3, we describe the modeling approach based on overparameterized vectors. In Section 4, we introduce the proposed identification scheme, and we give some theoretical background in Section 5. Numerical experiments are illustrated in Section 6. Some conclusions end the paper.

## 2 Problem formulation

We consider a stable single input single output discrete-time system described by the following time-domain relations (see Figure 1)

$$\begin{aligned} w_t &= f(u_t) \\ y_t &= \sum_{k=1}^{\infty} g_k w_{t-k} + e_t. \end{aligned} \quad (1)$$

In the above equation,  $f(\cdot)$  represents a (static) nonlinear function transforming the measurable input  $u_t$  into the unavailable signal  $w_t$ , which in turn feeds a strictly causal stable LTI system, described by the impulse response  $g_t$ . The output measurements of the system  $y_t$  are corrupted by white Gaussian noise, denoted by  $e_t$ , which has unknown variance  $\sigma^2$ . Following a standard approach in Hammerstein system identification (see e.g. [2]), we assume that  $f(\cdot)$  can be modeled as a combination of  $p$  known basis functions  $\{\phi_i\}_{i=1}^p$ , namely

$$w_t = f(u_t) = \sum_{i=1}^p c_i \phi_i(u_t), \quad (2)$$

where the coefficients  $c_i$  are unknown.

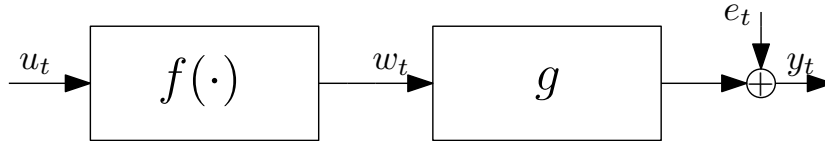


Figure 1: Block scheme of the Hammerstein system.

We assume that  $N$  input-output samples are collected, and denote them by  $\{u_t\}_{t=0}^{N-1}$ ,  $\{y_t\}_{t=1}^N$ . For notational convenience, we also assume null initial conditions. Then, the system identification problem we discuss in this paper is the problem of estimating  $n$  samples of the impulse response, say  $\{g_t\}_{t=1}^n$  (where  $n$  is large enough to capture the system dynamics), as well as the  $p$  coefficients  $\{c_i\}_{i=1}^p$  characterizing the static nonlinearity  $f(\cdot)$ .

### 2.1 Non-uniqueness of the identified system

It is well-known (see e.g. [3]) that the two components of a Hammerstein system can be determined up to a scaling factor. In fact, for any  $\alpha \in \mathbb{R}$ , every pair

$(\alpha g_t, \frac{1}{\alpha} f(\cdot))$ , describes the input-output relation equally well. As suggested in [3], we will circumvent this non-uniqueness issue by introducing the following assumption:

**Assumption 1.** *The impulse response has unitary  $\ell_2$  gain, i.e.  $\|g\|_2 = 1$ , and the sign of its first non-zero element is positive.*

## 2.2 Notation and preliminaries

Given a sequence of scalars  $\{a_t\}_{t=1}^m$ , we denote by  $a$  its vector representation, i.e.

$$a = \begin{bmatrix} a_1 \\ \vdots \\ a_m \end{bmatrix} \in \mathbb{R}^m.$$

We reserve the symbol  $\otimes$  to indicate the Kronecker product of two matrices (or vectors). We will make use of the bilinear property

$$(A \otimes B)(C \otimes D) = AC \otimes BD,$$

where  $A$ ,  $B$ ,  $C$ , and  $D$  have proper dimensions. Denoting by  $\text{vec}(A)$  the column-wise vectorization of a matrix  $A$ , we recall that, for any two vectors  $a$  and  $b$ ,  $\text{vec}(ab^T) = b \otimes a$ . Given a vector  $a \in \mathbb{R}^{np}$ , we introduce its  $n \times p$  reshape as

$$\mathcal{R}_{n,p}(a) := \begin{bmatrix} a_1 & \cdots & a_{n(p-1)+1} \\ \vdots & & \vdots \\ a_n & \cdots & a_{np} \end{bmatrix} \in \mathbb{R}^{n \times p}.$$

Given  $a \in \mathbb{R}^m$ , The symbol  $\mathbf{T}_n(a)$  denotes the  $m \times n$  Toeplitz matrix whose entries are elements of  $a$ , namely

$$\mathbf{T}_n(a) = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ a_2 & a_1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ a_{m-1} & a_{m-2} & \cdots & a_{m-n} & 0 \\ a_m & a_{m-1} & \cdots & \cdots & a_{m-n+1} \end{bmatrix} \in \mathbb{R}^{m \times n}. \quad (3)$$

Let

$$S = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ & & & 0 \\ & I_{m-1} & & \vdots \\ & & & 0 \end{bmatrix} \in \mathbb{R}^{m \times p} \quad (4)$$

and

$$P = [I \quad S \quad S^2 \quad \cdots \quad S^{n-1}]. \quad (5)$$

We have the following result, which will be used throughout the paper.

**Lemma 1.** *Let  $a \in \mathbb{R}^m$  and  $\mathbf{T}_n(a)$  be as in (3). Then*

$$P(I \otimes a) = \mathbf{T}_n(a), \quad (6)$$

*Proof.* Note that

$$\begin{aligned} \mathbf{T}_n(a) &= [a \quad Sa \quad S^2a \quad \cdots \quad S^{n-1}a] \\ &= [I \quad S \quad S^2 \quad \cdots \quad S^{n-1}] \begin{bmatrix} a \\ \vdots \\ a \end{bmatrix} \\ &= P(I \otimes a), \end{aligned} \quad (7)$$

which proves the statement.  $\square$

Based on the equality stated by Lemma 1, we extend the Toeplitz notation to matrices, that is, given  $A \in \mathbb{R}^{m \times p}$  we write

$$\mathbf{T}_n(A) = P(I_n \otimes A) \in \mathbb{R}^{m \times np}. \quad (8)$$

### 3 Identification via overparameterized models

In this paper, we deal with overparameterized approaches to Hammerstein system identification. To this end, we construct the matrices

$$F \triangleq \begin{bmatrix} \phi_1(u_0) & \cdots & \phi_p(u_0) \\ \vdots & \vdots & \vdots \\ \phi_1(u_{N-1}) & \cdots & \phi_p(u_{N-1}) \end{bmatrix}, \quad (9)$$

and

$$\Phi \triangleq \mathbf{T}_n(F) \in \mathbb{R}^{N \times np}. \quad (10)$$

Then, we can express the Hammerstein system dynamics problem by means of the linear regression model (see also [2])

$$y = \Phi \vartheta + e, \quad (11)$$

where

$$\vartheta = g \otimes c \in \mathbb{R}^{np}. \quad (12)$$

This vector contains the  $n + p$  unknown parameters of the Hammerstein model. Thus, it constitutes an overparameterization with respect to the original parameters  $c$  and  $g$ . A desirable property of any estimate of  $\vartheta$  is that it should be expressible as (12), namely as a Kronecker product of two vectors. We formalize this concept in the following definition.

**Definition 1.** *Let  $\vartheta \in \mathbb{R}^{np}$ . We say that  $\vartheta$  is a Kronecker overparameterized (KOP) vector if there exist  $g \in \mathbb{R}^n$  and  $c \in \mathbb{R}^p$  such that (12) holds.*

The following lemma gives a property of KOP vectors.

**Lemma 2.** *Let  $\vartheta \in \mathbb{R}^{np}$  be a KOP vector. Then  $\mathcal{R}_{n,p}(\vartheta) = cg^T$  and thus  $\mathcal{R}_{n,p}(\vartheta)$  is a rank-one matrix.*

*Proof.* Follows from the identity  $\text{vec}(cg^T) = g \otimes c$ .  $\square$

Under Assumption 1, an  $np$ -dimensional KOP vector  $\vartheta$  can be uniquely decomposed into the  $n$ - and  $p$ - dimensional vectors  $g$  and  $c$ .

**Proposition 1.** *Let  $\vartheta \in \mathbb{R}^{np}$  be a KOP vector. Let  $\tilde{g}$  be the  $i$ th row and  $\tilde{c}$  and the  $j$ th column of  $\mathcal{R}_{n,p}(\vartheta)$ , define*

$$g = \frac{\tilde{g}}{\|\tilde{g}\|} \text{sign}(\tilde{g}_1), \quad c = \frac{\tilde{c}}{\tilde{g}_j} \|\tilde{g}\| \text{sign}(\tilde{g}_1). \quad (13)$$

*Then  $\vartheta = g \otimes c$ ,  $\|g\| = 1$  and  $g_1 > 0$ .*

*Proof.* From (13) we have that  $c_i g_j = \tilde{c}_i$ , so  $c_i g_j$  is the  $i, j$ th element of  $\mathcal{R}_{n,p}(\vartheta)^T$  hence  $\mathcal{R}_{n,p}(\vartheta)^T = cg^T$ . In addition

$$\|g\| = \left\| \frac{\tilde{g}}{\|\tilde{g}\|} \text{sign}(\tilde{g}_1) \right\| = \frac{\|\tilde{g}\|}{\|\tilde{g}\|} |\text{sign}(\tilde{g}_1)| = 1, \quad (14)$$

and

$$g_1 = \frac{\tilde{g}_1}{\|\tilde{g}\|} \text{sign}(\tilde{g}_1) = \frac{|\tilde{g}_1|}{\|\tilde{g}\|} > 0, \quad (15)$$

which completes the proof.  $\square$

### 3.1 A review of an overparameterized method for Hammerstein system identification

In this section we review the identification procedure proposed in [2], which constitutes the starting point of our regularized kernel-based method. Given the model (11), consistent estimates of  $c$  and  $g$  can be obtained with the following steps (see [2] for details about consistency). First, we compute the least-squares estimate

$$\hat{\vartheta}^{\text{LS}} = (\Phi^T \Phi)^{-1} \Phi^T y. \quad (16)$$

Then, since we know that  $\vartheta$  is a KOP vector, that is, the reshaping of  $\vartheta$  into an  $n \times p$  matrix must be rank-one (Lemma 2), we approximate  $\hat{\vartheta}^{\text{LS}}$  to a KOP vector by approximating  $\mathcal{R}_{n,p}(\hat{\vartheta}^{\text{LS}})$  to a rank-one matrix. This can be done by solving the problem

$$\begin{aligned} & \text{minimize} && \|X - \mathcal{R}_{n,p}(\hat{\vartheta}^{\text{LS}})\|_F \\ & \text{s.t.} && \text{rank } X = 1, \end{aligned} \quad (17)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm. Expressing  $\mathcal{R}_{n,p}(\hat{\vartheta}^{\text{LS}})$  by means of its singular value decomposition, i.e.

$$\begin{aligned}\mathcal{R}_{n,p}(\hat{\vartheta}^{\text{LS}}) &= USV^T \\ &= [u^1 \ \dots \ u^p] \text{diag}\{s_1, \dots, s_p\} [v^1 \ \dots \ v^p]^T,\end{aligned}\tag{18}$$

we find that the solution of (17) is  $X = u^1 s_1 v^{1T}$ . Then  $\hat{g} = v^1 \text{sign}(v_1^1)$  (since we have assumed  $\|g\|_2 = 1$ ) and  $\hat{c} = s_1 u^1 \text{sign}(v_1^1)$ .

Note that, since in general  $\vartheta^{\text{LS}}$  is not a KOP vector, generally  $s_2, \dots, s_p > 0$  and the truncation required by the approximation (17) introduces a bias in the estimates  $\hat{g}$  and  $\hat{c}$  that degrades performance (see [14]). Another drawback of this method is that it requires the least-squares estimate of the possibly high-dimensional vector  $\hat{\vartheta}^{\text{LS}}$ . Hence, despite its consistency property, the procedure can suffer from high variance in the estimates when  $N$  is not (very) large.

## 4 A regularized overparameterization method for Hammerstein system identification

In the previous section, we have seen that the estimator proposed in [2] suffers from high variance and from a bias that degrades performance. To control the variance of the estimate, we can use regularization (for a full treatment, see [5], [13]); this means we have to select some properties we want to enforce on the estimated vector. As we have pointed out in the previous section, a vector  $\hat{\vartheta}$  is a good candidate estimate of the unknown vector if it satisfies the following properties:

1.  $\hat{\vartheta}$  is a KOP vector, so that it can be decomposed as in (12);
2. The mean square error of  $\hat{\vartheta}$  is low, so that the estimated values  $\hat{g}$  and  $\hat{c}$  are close to the true values.

A natural approach to incorporate (at least) the second property is based on regularization or, equivalently, on the Gaussian regression framework [29]. Thus, we model  $\vartheta$  as a Gaussian random vector, namely

$$\vartheta \sim \mathcal{N}(0, H(\rho)).\tag{19}$$

where the covariance matrix (also called a *kernel*)  $H(\rho)$  is parameterized by the vector  $\rho$ . The structure of  $H(\rho)$  determines the properties of the realizations from (19) and, consequently, of the estimates of  $\vartheta$ . In the next subsection, we focus on designing a kernel suitable for Hammerstein system identification that incorporates also the first property.

### 4.1 The KOP kernel

We first recall the kernel-based identification approach for LTI systems proposed in [20], [21], and we model also  $g$  as a realization of a zero-mean  $n$ -dimensional

Gaussian process. Then we have

$$g \sim \mathcal{N}(0, K_\beta), \quad (20)$$

where the kernel  $K_\beta$  corresponds to the so-called *first-order stable spline kernel* (or *TC kernel* in [7]). It is defined as

$$\{K_\beta\}_{i,j} \triangleq \beta^{\max(i,j)}, \quad (21)$$

where the hyperparameter  $\beta$  is a scalar in the interval  $[0, 1)$ . The choice of this kernel is motivated by the fact that it promotes BIBO stable and smooth realizations. The decay velocity of these realizations is regulated by  $\beta$ . Typical formulations of the stable spline kernel (see e.g. [23]) include a scaling factor multiplying the kernel, in order to capture the amplitude of the unknown impulse response. Here such an hyperparameter is redundant, as we are working under Assumption 1.

To reconcile (20) with (19), we need to ensure that the transformation  $g \otimes c$  is a Gaussian vector, when  $g$  is Gaussian. This is possible if  $c$  is a (deterministic)  $p$ -dimensional vector. In this case  $\vartheta$  is an  $np$ -dimensional Gaussian random vector with covariance matrix

$$H(\rho) = \mathbb{E}[\vartheta \vartheta^T] = \mathbb{E}[(g \otimes c)(g \otimes c)^T] = K_\beta \otimes cc^T, \quad (22)$$

which is parameterized by the vector  $\rho = [\beta \ c^T]^T$ . In this way, we have defined a new kernel for system identification based on overparameterized vector regression. We formalize this in the following definition.

**Definition 2.** We define the Kronecker overparameterized (KOP) kernel as

$$H(\rho) \triangleq K_\beta \otimes cc^T, \quad \rho = [\beta \ c^T]^T, \quad (23)$$

where  $K_\beta$  is as in (21).

Note that  $H(\rho)$  is rank-deficient, its rank being equal to  $n$ . Rank-deficient kernels for system identification have also been studied in [8].

## 4.2 Estimation of the overparameterized vector $\vartheta$

We now derive the estimation procedure for the vector  $\vartheta$ . Recalling that the noise distribution is Gaussian and given the Gaussian description of  $\vartheta$  (19), the joint distribution of  $y$  and  $\vartheta$  is Gaussian. Hence, we can write

$$p\left(\begin{bmatrix} y \\ \vartheta \end{bmatrix}; \rho, \sigma^2\right) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_y & \Sigma_{y\vartheta} \\ \Sigma_{\vartheta y} & H(\rho) \end{bmatrix}\right), \quad (24)$$

where  $\Sigma_{y\vartheta} = \Sigma_{\vartheta y}^T = \Phi H(\rho)$  and

$$\Sigma_y = \Phi H(\rho) \Phi^T + \sigma^2 I. \quad (25)$$



In (24) we have highlighted the dependence of the joint distribution on the vector  $\rho$  and the noise variance  $\sigma^2$ . Assume these quantities are given; then the minimum mean square error estimate of  $\vartheta$  can be computed as (see e.g. [1])

$$\begin{aligned}\hat{\vartheta} &= \mathbb{E}[\vartheta|y; \rho, \sigma^2] \\ &= H(\rho)\Phi^T\Sigma_y^{-1}y.\end{aligned}\tag{26}$$

To be able to compute (24) we first need to determine  $\rho$  and  $\sigma^2$ . This can be done by maximizing the ML of the output data (see e.g. [19]). Then we have

$$\begin{aligned}\hat{\rho}, \hat{\sigma}^2 &= \arg \max p(y; \rho, \sigma^2) \\ &= \arg \min_{\rho, \sigma^2} \left[ \log \det \Sigma_y + y^T \Sigma_y^{-1} y \right].\end{aligned}\tag{27}$$

The resulting estimation procedure for  $\vartheta$  can be summarized by the following two steps.

1. Solve (27) to obtain  $\hat{\rho}, \hat{\sigma}^2$ .
2. Compute (26) using the estimated parameters.

Having obtained  $\hat{\vartheta}$ , it remains to establish how to decompose it in order to obtain the estimates  $\hat{g}$  and  $\hat{c}$ . In the next section we shall see that, using the proposed approach, such an operation becomes natural.

## 5 Properties of the estimated overparameterized vector

In this section, we analyze some properties of the regularized overparameterization estimate of  $\vartheta$ . In particular, we show that the estimates produced by (26) are KOP vectors. Then, we show that this procedure leads to exactly the same estimator as the one we proposed in [25]; where the coefficients of the nonlinearity were considered as model parameters and not included among the kernel hyperparameters.

To further specify the equivalence, we first briefly review the Hammerstein system identification approach proposed in [25] which is based on a different Gaussian process assumption.

### 5.1 A review of the method proposed in [25]

Let  $W = \mathbf{T}_n(w) = \mathbf{T}_n(Fc)$ . Then we can model the measurements with the linear relation

$$y = Wg + e.\tag{28}$$

Modeling  $g$  as a Gaussian random vector with covariance given by the stable spline kernel (21), we notice that a joint Gaussian description holds between  $y$

and  $g$ . Hence we can write

$$p\left(\begin{bmatrix} y \\ g \end{bmatrix}; c, \beta, \sigma^2\right) \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{y,2} & \Sigma_{yg} \\ \Sigma_{gy} & H(\rho) \end{bmatrix}\right), \quad (29)$$

where  $\Sigma_{yg} = WK_\beta$  and  $\Sigma_{y,2} = WK_\beta W^T + \sigma^2 I$ . Note that (29) depends on the parameters  $c$ ,  $\beta$  and  $\sigma^2$ . These parameters are estimated via ML maximization, that is by solving

$$\hat{c}, \hat{\beta}, \hat{\sigma}^2 = \arg \min_{c, \beta, \sigma^2} \left[ \log \det \Sigma_{y,2} + y^T \Sigma_{y,2}^{-1} y \right]. \quad (30)$$

The minimum mean square estimate of  $g$  is then computed as

$$\hat{g} = \mathbb{E}[g|y, \hat{c}, \hat{\beta}, \hat{\sigma}^2] = K_\beta W^T \Sigma_{y,2}^{-1} y. \quad (31)$$

In the next section, we point out the strong connection between (31) and the estimate (26), produced by the KOP kernel-based regression approach.

## 5.2 The estimate (26) is a KOP vector

In this section we prove that, when the KOP kernel-based method is used to estimate (26), the resulting estimates can be decomposed as Kronecker products of lower-dimensional vectors and thus they are KOP vectors. Before arriving at this result we show the equivalence between the output measurement models (11) and (28).

**Lemma 3.** *Let  $W = \mathbf{T}_n(Fc)$  and  $\Phi$  as in (10). Then*

$$\Phi H(\rho) \Phi^T = WK_\beta W, \quad (32)$$

where  $H(\rho)$  and  $K_\beta$  are the KOP and the stable spline kernels.

*Proof.* Recalling the bilinear property of the Kronecker product and Lemma 1, we see that

$$\begin{aligned} \Phi H(\rho) \Phi^T &= P[I \otimes F] H(\rho) [I \otimes F^T] P^T \\ &= P[I \otimes F] [K_\beta \otimes cc^T] [I \otimes F^T] P^T \\ &= P[I \otimes F] [I \otimes c] [K_\beta \otimes 1] [I \otimes c^T] [I \otimes F^T] P^T \\ &= P[I \otimes Fc] [K_\beta \otimes 1] [I \otimes c^T F^T] P^T \\ &= \mathbf{T}_n(Fc) K_\beta \mathbf{T}_n(Fc)^T = WK_\beta W^T, \end{aligned}$$

which proves the result.  $\square$

**Theorem 1.** *Consider the output measurement models (11) and (28). Then:*

1. *The marginal likelihoods of  $y$  obtained from the two models are equivalent;*
2. *The parameter estimates obtained from (27) and (30) are the same.*

*Proof.* Let

$$p_1(y; \rho, \sigma^2) = \int p(y, \vartheta; \rho, \sigma^2) d\vartheta \quad (33)$$

$$p_2(y; c, \beta, \sigma^2) = \int p(y, g; c, \beta, \sigma^2) dg \quad (34)$$

be the marginal likelihoods of the two models. We find that

$$\begin{aligned} p_1(y; \rho, \sigma^2) &= \mathcal{N}(0, \Sigma_y), \\ p_2(y; c, \beta, \sigma^2) &= \mathcal{N}(0, \Sigma_{y,2}). \end{aligned}$$

Using Lemma 3, we have that  $\Sigma_y = \Sigma_{y,2}$ , hence  $p_1$  and  $p_2$  are equivalent. The same promptly holds for their ML maximizers  $\hat{\rho} = [\hat{\beta} \ \hat{c}^T]^T$  and  $\hat{\sigma}^2$ .  $\square$

We are now in the position to prove that the estimate  $\hat{\vartheta}$  is a KOP vector

**Theorem 2.** *Assume that  $\rho$  and  $\sigma^2$  are estimated using the ML approach (27) (or, equivalently, (30)). Then, the minimum variance estimate of  $\vartheta$  in (26) is such that*

$$\hat{\vartheta} = \hat{g} \otimes \hat{c}, \quad (35)$$

where  $\hat{g}$  is the minimum variance estimate of  $g$  in (31) and  $\hat{c}$  is the ML estimate of  $c$ .

*Proof.* Using (26) and recalling the bilinear property of the Kronecker product and Lemma 1, we have

$$\begin{aligned} \hat{\vartheta} &= \Sigma_{\vartheta y} \Sigma_y^{-1} y = H(\hat{\rho}) \Phi^T \Sigma_y^{-1} y \\ &= [K_{\hat{\beta}} \otimes \hat{c} \hat{c}^T] [I \otimes F^T] P^T \Sigma_y^{-1} y \\ &= [K_{\hat{\beta}} \otimes \hat{c}] [I \otimes \hat{c}^T F^T] P^T \Sigma_y^{-1} y \\ &= [K_{\hat{\beta}} \otimes \hat{c}] W^T \Sigma_y^{-1} y \\ &= [K_{\hat{\beta}} \otimes \hat{c}] [W^T \Sigma_y^{-1} y \otimes 1] \\ &= [K_{\hat{\beta}} W^T \Sigma_y^{-1} y \otimes \hat{c}]. \end{aligned} \quad (36)$$

From Theorem 1 we know that  $\Sigma_y = \Sigma_{y,2}$ ; thus, recalling (31) we have

$$K_{\hat{\beta}} W^T \Sigma_y^{-1} y = \hat{g}, \quad (37)$$

so that (35) is obtained.  $\square$

**Corollary 1.** *The estimate  $\hat{\vartheta}$  given in (26) is a KOP vector and  $\mathcal{R}_{n,p}(\hat{\vartheta})$  is rank-one.*

*Proof.* Since from (35) we have  $\hat{\vartheta} = \hat{g} \otimes \hat{c}$ ,  $\hat{\vartheta}$  is a KOP vector. The second part of the statement follows directly from Lemma 2.  $\square$

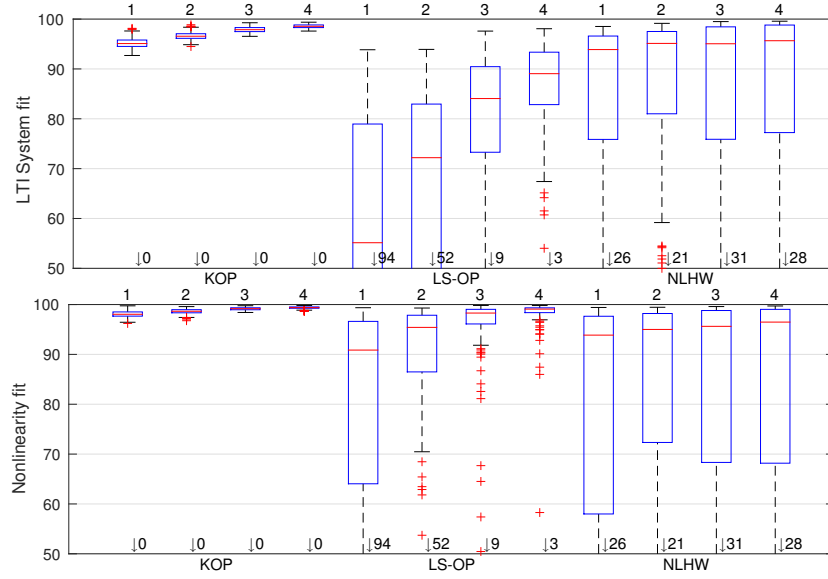


Figure 2: Results of the Monte Carlo experiments for different SNR. Top: Fit in percent of the linear system impulse response. Bottom: Fit in percent of the nonlinear transformation.

We can make an interesting observation, that further links the KOP estimate to our previous kernel based estimator:

**Corollary 2.** *The estimates of the nonlinearity coefficients  $\hat{c}$ , found maximizing (27) and those resulting from decomposing  $\hat{\vartheta}$  as in (35) are the same.*

*Proof.* Follows directly from (36) and Theorem 1.  $\square$

We have established that the estimate  $\hat{\vartheta}$  produced using the procedure detailed in Section 4 is a KOP vector. So, the estimates of the impulse response  $\hat{g}$  and the nonlinearity coefficients  $\hat{c}$  can be retrieved using (13). The whole procedure is summarized in Algorithm 1.

---

**Algorithm 1:** KOP kernel-based Hammerstein system identification

---

Input:  $\{y_t\}_{t=1}^N, \{u_t\}_{t=0}^{N-1}$   
Output:  $\{\hat{g}_t\}_{t=1}^n, \{\hat{c}_i\}_{i=1}^p$

1. Obtain  $\hat{\rho}, \hat{\sigma}^2$  solving (27)
  2. Estimate  $\hat{\vartheta}$  using (26)
  3. Find  $\hat{g}$ , normalizing  $\hat{\vartheta}$  by  $\hat{c}$  (Proposition 1).
-

The result of the outlined regularization procedure applied to the overparameterized vector, with a suitable rank deficient kernel, yields estimates that are equivalent to the ones provided by the procedure outlined in [25].

## 6 Numerical Experiments

We evaluate the proposed algorithm with numerical simulations of Hammerstein systems. We set up 4 experiments in order to test different experimental conditions. The experiments consist of 200 independent Monte Carlo runs each. At every Monte Carlo run, we generate random data and Hammerstein systems, according to the following specifics.

- The linear subsystem model is of output error type:

$$y(t) = \frac{B(q)}{A(q)} f(u(t)) + e(t), \quad (38)$$

generated by picking 4 poles and 4 zeros at random. The poles and zeros were sampled in conjugate pairs  $(ae^{j\omega}, ae^{-j\omega})$  with  $a$  uniform in  $[0.5, 0.95]$  and  $\omega$  uniform in  $[0, \pi]$ .

- The input nonlinearity is a polynomial of fourth order. It is a linear combination of Legendre polynomial basis functions, defined as

$$P_i(u) = \frac{1}{2^i i!} \frac{\partial^i}{\partial u^i} \left[ (u^2 - 1)^i \right], \quad (39)$$

where  $i = 0, \dots, 4$ . The coefficients  $c$  are chosen uniformly in  $[-1, 1]$ .

- The input to the system is Gaussian white noise with unit variance.
- The experimental data consists in  $N = 1000$  pairs of input-output samples, simulated from zero initial conditions.
- The measurement noise  $e(t)$  is Gaussian and white. Its variance is a fraction of the noiseless output variance, i.e.

$$\sigma^2 = \frac{\text{Var } Wg}{SNR}, \quad (40)$$

where  $SNR$  depends on the experiment.

Every experiment is carried out in a different signal to noise ratio (SNR) condition, see Table I.

We aim at estimating  $n = 30$  samples of the system impulse response of the LTI systems (which are such that  $\|g\|_2 = 1$  and with the sign of the first sample positive) and the  $p = 5$  coefficients of the nonlinear block. We test the following estimation methods.

Experiment #	1	2	3	4
SNR	10	20	50	100

Table 1: SNR considered in the experiments

- *KOP*: This is the method described in this paper. The ML optimization problem is solved using the function `fminsearch` available in MATLAB. The search was initialized with the elements of  $c$  uniformly sampled in  $[0, 1]$ ,  $\beta_0 = 0.5$ , and  $\sigma$  equal to the sample variance of the residuals of  $\hat{v}^{\text{LS}}$ .
- *LS-OP*: This estimator implements the least-squares overparameterization-based method proposed in [2] and briefly reviewed in Section 3.1. Note that, under the working experimental conditions, this method has to perform a least-squares estimate of a 150-dimensional vector.
- *NLHW*: This is the MATLAB function `nlhw` that uses the prediction error method to identify the linear block in the system (see [17] for details). To get the best performance from this method, we equip it with an oracle that knows the true order of the LTI system generating the measurements and knows the order of the polynomial input nonlinearity.

Note that all the methods have available the same amount of prior information, namely the orders of the input polynomial. The knowledge of the order of the linear block is known only to NLHW, which makes use of a parametric description of the linear system. Furthermore, we note that, due to the Gaussianity of the noise, the least-squares procedure in LS-OP is optimal in the Gauss-Markov sense.

We assess the accuracy of the estimated models using two performance indices. The first is the fit of the system impulse response, defined as

$$FIT_{g,i} \triangleq 100 \left( 1 - \frac{\|g_i - \hat{g}_i\|_2}{\|g_i - \bar{g}_i\|_2} \right), \quad (41)$$

where  $g_i$  is the system generated at the  $i$ -th run of each experiment,  $\hat{g}_i$  its estimate and  $\bar{g}_i$  its mean. The second is the fit of the static nonlinear function, given by

$$FIT_{f,i} \triangleq 100 \left( 1 - \frac{\|f_i(u) - \hat{f}_i(u)\|_2}{\|f_i(u) - \bar{f}_i(u)\|_2} \right). \quad (42)$$

Figure 1 shows the results of the outcomes of the 4 experiments. The box plots compare the results of KOP, LS-OP and NLHW for the considered signal to noise ratios. We can see that, for high SNR, all the estimators perform well, especially in identifying the nonlinearity coefficients. For lower SNR, however, the proposed method KOP performs substantially better than the others. This is mainly because of the regularizing effect of the KOP kernel that reduces

the variance of the estimates. Notice also that the proposed approach enforces the rank deficiency in the reshaped version of  $\hat{\vartheta}$ , so it circumvents the errors introduced by the rank-one approximation made by LS-OP. The main drawback of NLHW is that it relies on a high dimensional nonlinear optimization, as it needs to estimate all the parameters in the model. The proposed method is instead nonparametric, and does not rely on the knowledge of the order of the LTI system.

## 7 Conclusions

Regularization is an effective technique to control the variance of least squares estimates. In this paper we have studied how to improve popular overparameterization methods for Hammerstein system identification using Gaussian process regression with a suitable prior. To this end, starting from the stable spline kernel, we have introduced the KOP kernel, which we believe to be a novel concept in Hammerstein system identification. Using the KOP kernel, we have designed a regularized least-squares estimator which provides an estimate of the overparameterized vector. The impulse response of the LTI system and the coefficients of the static nonlinearity are then retrieved by suitably decomposing the estimated vector. In contrast with the original overparameterization method, this decomposition involves no approximation. An important contribution is showing that this procedure estimate is equivalent to our recently proposed kernel-based method [25]. Using simulations, we have shown that the proposed method compares very favorably with the current state-of-the-art algorithms for Hammerstein system identification.

The introduction of the KOP kernel possibly opens up for new effective system identification methods based on the combination of overparameterized vectors and regularization techniques. In fact, we believe that Hammerstein system identification is not the only problem where KOP kernels could find application. Another possible extension of the proposed method is the design of new kernels merging a kernel for the static nonlinearity and the stable spline kernel. The main issue with this approach is that, at least theoretically, the Gaussian description of the resulting overparameterization vector would be lost.

## References

- [1] B. D. O. Anderson and J. B. Moore. *Optimal filtering*. Courier Corporation, 2012.
- [2] E. W. Bai. “An optimal two-stage identification algorithm for Hammerstein–Wiener nonlinear systems”. *Automatica* 34.3 (1998), pp. 333–338.
- [3] E. W. Bai and D. Li. “Convergence of the iterative Hammerstein system identification algorithm”. *IEEE Trans. Autom. Control* 49.11 (2004), pp. 1929–1940.

- [4] E. W. Bai et al. “Identification of a modified Wiener–Hammerstein system and its application in electrically stimulated paralyzed skeletal muscle modeling”. *Automatica* 45.3 (2009), pp. 736–743.
- [5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [6] G. Bottegal and G. Pillonetto. “Regularized spectrum estimation using stable spline kernels”. *Automatica* 49.11 (2013), pp. 3199–3209.
- [7] T. Chen, H. Ohlsson, and L. Ljung. “On the estimation of transfer functions, regularizations and Gaussian processes—Revisited”. *Automatica* 48.8 (2012), pp. 1525–1535.
- [8] T. Chen et al. “Rank-1 kernels for regularized system identification”. *52nd IEEE Conference on Decision and Control*. Institute of Electrical & Electronics Engineers (IEEE), 2013.
- [9] I. Goethals et al. “Subspace identification of Hammerstein systems using least squares support vector machines”. *IEEE Trans. Autom. Control* 50.10 (2005), pp. 1509–1519.
- [10] W. Greblicki. “Stochastic approximation in nonparametric identification of Hammerstein systems”. *IEEE Transactions on Automatic Control* 47.11 (2002), pp. 1800–1810.
- [11] W. Greblicki and M. Pawlak. “Identification of discrete Hammerstein systems using kernel regression estimates”. *IEEE Trans. Autom. Control* 31.1 (1986), pp. 74–77.
- [12] Y. Han and R. De Callafon. “Hammerstein system identification using nuclear norm minimization”. *Automatica* 48.9 (2012), pp. 2189–2193.
- [13] T. Hastie et al. *The elements of statistical learning: data mining, inference and prediction*. Springer-Verlag New York Inc., 2009. ISBN: 0387848576.
- [14] H Hjalmarsson and J Schoukens. “On direct identification of physical parameters in non-linear models”. *Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS), Stuttgart, Germany*. 2004.
- [15] I. Hunter and M. Korenberg. “The identification of nonlinear biological systems: Wiener and Hammerstein cascade models”. *Biol. Cybern.* 55.2-3 (1986), pp. 135–144.
- [16] L. Ljung. *System Identification, Theory for the User*. Prentice Hall, 1999.
- [17] L. Ljung et al. “Developments in The MathWorks System Identification Toolbox”. *IFAC Proceedings Volumes* 42.10 (2009), pp. 522–527.
- [18] J. Maritz and T. Lwin. *Empirical bayes methods*. Chapman and Hall London, 1989.
- [19] G. Pillonetto and A. Chiuso. “Tuning complexity in kernel-based linear system identification: The robustness of the marginal likelihood estimator”. *Proc. European Control Conf. (ECC)*. 2014, pp. 2386–2391.



- [20] G. Pillonetto, A. Chiuso, and G. De Nicolao. “Regularized estimation of sums of exponentials in spaces generated by stable spline kernels”. *Proc. Amer. Control Conf. (ACC)*. 2010, pp. 498–503.
- [21] G. Pillonetto, A. Chiuso, and G. De Nicolao. “Prediction error identification of linear systems: a nonparametric Gaussian regression approach”. *Automatica* 47.2 (2011), pp. 291–305.
- [22] G. Pillonetto and G. De Nicolao. “A new kernel-based approach for linear system identification”. *Automatica* 46.1 (2010), pp. 81–93.
- [23] G. Pillonetto et al. “Kernel methods in system identification, machine learning and function estimation: A survey”. *Automatica* 50.3 (2014), pp. 657–682.
- [24] S. Rangan, G. Wolodkin, and K. Poolla. “New results for Hammerstein system identification”. *Proc. IEEE Conf. Decis. Control (CDC)*. Vol. 1. IEEE. 1995, pp. 697–702.
- [25] R. S. Risuleo, G. Bottegal, and H. Hjalmarsson. “A kernel-based approach to Hammerstein system identification”. *Proc. IFAC Symp. System Identification (SYSID)*. Vol. 48. 28. 2015, pp. 1011–1016.
- [26] M. Schoukens, R. Pintelon, and Y. Rolain. “Parametric Identification of Parallel Hammerstein Systems”. *IEEE Trans. Instrum. Meas.* 60.12 (2011), pp. 3931–3938.
- [27] G. Wahba. *Spline models for observational data*. Vol. 59. SIAM, 1990. ISBN: 0898712440.
- [28] D. Westwick and R. Kearney. “Separable least squares identification of nonlinear Hammerstein models: Application to stretch reflex dynamics”. *Ann. Biomed. Eng.* 29.8 (2001), pp. 707–718.
- [29] C. Williams and C. Rasmussen. *Gaussian processes for machine learning*. 2006.